Machine Learning for Synthetic Data Generation: a Review

Year: 2023 | Citations: 217 | Authors: Ying-Cheng Lu, Huazheng Wang, Wenqi Wei

Abstract

Machine learning heavily relies on data, but real-world applications often encounter various data-related issues. These include data of poor quality, insufficient data points leading to under-fitting of machine learning models, and difficulties in data access due to concerns surrounding privacy, safety, and regulations. In light of these challenges, the concept of synthetic data generation emerges as a promising alternative that allows for data sharing and utilization in ways that real-world data cannot facilitate. This paper presents a comprehensive systematic review of existing studies that employ machine learning models for the purpose of generating synthetic data. The review encompasses various perspectives, starting with the applications of synthetic data generation, spanning computer vision, speech, natural language processing, healthcare, and business domains. Additionally, it explores different machine learning methods, with particular emphasis on neural network architectures and deep generative models. The paper also addresses the crucial aspects of privacy and fairness concerns related to synthetic data generation. Furthermore, this study identifies the challenges and opportunities prevalent in this emerging field, shedding light on the potential avenues for future research. By delving into the intricacies of synthetic data generation, this paper aims to contribute to the advancement of knowledge and inspire further exploration in synthetic data generation.